Le contrôle humain n'est pas extérieur au système

Le système transforme celui qui le contrôle. Sur la capacité humaine indépendante minimale, et sur l'absence d'instrument pour la mesurer.

1. Le paradoxe

Le 8 septembre 2026, le *Wall Street Journal* publiait en exclusivité la démission de Jacob Coxon, chercheur en pré-entraînement passé par OpenAI puis Anthropic, qui reprochait à ces laboratoires de foncer vers une superintelligence auto-améliorante en jouant nos vies (Ramkumar A., « Anthropic Researcher Quits Over "Out-of-Control" AI Fears », *WSJ*, 8 septembre 2026). Il déclarait au journal que le monde s'orientait vers les plus agressifs de ces scénarios, où dès la fin de l'année prochaine les choses pourraient être hors de contrôle, et écrivait par ailleurs que ceux qui construisent ces systèmes pensent qu'ils pourraient tous nous tuer d'ici la fin de la décennie.

Il faut prendre cette alerte au sérieux et noter ce qu'elle contient. Chacune de ses affirmations est indexée sur un futur : la fin de l'année prochaine, la fin de la décennie. Aucune ne porte sur un état mesuré aujourd'hui, et ce n'est pas un défaut de rigueur de sa part, c'est la nature de l'objet. *Un risque dont la probabilité reste indéterminée peut parfaitement justifier des travaux de précaution, réduction d'incertitude, confinement, robustesse, exercices adverses ; il offre en revanche peu de prise à une comparaison quantitative de priorité.* Un impact potentiellement illimité ne fonde pas un rang, faute de comparateur. Il existe un autre risque, moins spectaculaire, déjà observable, et directement rattaché au dispositif réglementaire en vigueur.

Le voici. Le droit exige un contrôle humain sur les systèmes d'intelligence artificielle à haut risque. Ce faisant, il suppose implicitement que l'usage du système ne transforme pas la personne chargée de le contrôler. **Cette hypothèse est fausse, et c'est la thèse de cet article : le contrôle humain n'est pas une capacité extérieure appliquée au système, c'est une variable d'état couplée à lui. La gouvernance horizontale de l'IA ne définit aujourd'hui ni métrique explicite de cette variable, ni protocole standardisé permettant d'en mesurer la dépréciation et la reconstitution.**

Il faut être précis sur la nature du reproche, car sa version large serait fausse. Certains secteurs disposent déjà de mécanismes partiels qui en tiennent lieu : *proficiency testing*, évaluation périodique de compétence, double lecture, accord inter-observateurs, contrôle qualité clinique, re-certification. Ce qui manque n'est pas toute mesure, c'est la

mesure de la bonne chose au bon endroit. Ces dispositifs ont été construits pour surveiller la variabilité entre praticiens, pas la dérive d'un praticien exposé à une assistance algorithmique. Et le socle horizontal, celui qui s'applique à tous les systèmes à haut risque quel que soit le domaine, n'en contient aucun.

De même, dire que le droit traite le contrôle humain comme un paramètre n'est pas une affirmation sur la lettre du règlement, qui ne contient évidemment aucun modèle formel. C'est une affirmation sur son ontologie implicite. *Le droit le traite fonctionnellement comme un paramètre : il vérifie que la capacité est présente et n'en modélise jamais l'évolution sous l'effet de l'usage du système.* La distinction compte, parce que la lacune que ce texte cherche à établir est instrumentale et non normative.

2. Ce que l'on sait, ce que l'on ne sait pas, et pourquoi

L'état des preuves se lit en quatre niveaux, et il faut les tenir séparés. Les confondre est exactement ce qui permet à chaque camp de citer l'étude qui l'arrange.

Premier niveau, le signal.

En août 2025, *The Lancet Gastroenterology and Hepatology* publiait une étude que ses auteurs qualifient de découverte incidente (Budzyń K., Romańczyk M., Kitala D. et al., *Lancet Gastroenterol Hepatol* 2025;10(10):896-903). Quatre centres polonais engagés dans l'essai ACCEPT avaient déployé fin 2021 un système d'aide à la détection de polypes, les examens étant ensuite répartis entre coloscopies assistées et non assistées selon la date. Entre le 8 septembre 2021 et le 9 mars 2022, 1 443 patients ont subi une coloscopie **non assistée**, 795 avant l'introduction de l'outil et 648 après. Le taux de détection d'adénomes est passé de 28,4 % à 22,4 %, soit une différence absolue de six points (IC 95 % de -10,5 à -1,6 ; p = 0,0089), l'exposition à l'IA ressortant comme facteur indépendant en analyse multivariée avec un odds ratio de 0,69 (0,53-0,89). Dix-neuf endoscopistes, tous expérimentés. Une correction a depuis été publiée : l'indication de la coloscopie avait été omise des covariables, ce qui modifie les coefficients associés à la période postérieure à l'introduction de l'IA et à l'âge supérieur à soixante ans, sans que les auteurs et l'éditeur en tirent de changement d'interprétation.

Ce qui rend cette étude notable n'est pas son chiffre mais son objet. *La littérature sur l'assistance mesure la performance du couple humain-machine ; celle-ci mesure ce que devient la performance humaine quand l'assistance est retirée.* Le déplacement d'estimand est le véritable apport, et c'est un estimand que les protocoles d'évaluation ne comportent presque jamais.

Deuxième niveau, la tentative de réfutation.

Un essai pragmatique prospectif multicentrique sur registre, publié en 2026 dans *Endoscopy* (Pedersen et al., *Endoscopy* 2026;58(9):1003-1014), a mesuré le même phénomène avec un protocole nettement supérieur : treize endoscopistes, sept non expérimentés et six expérimentés, 5 013 coloscopies, et trois phases successives, avant exposition au système, pendant son usage, puis après son retrait. Critère principal, la proportion de coloscopies détectant au moins un polype d'au moins cinq millimètres.

Chez les non expérimentés, ce taux monte de 31,9 % (216 sur 678) en phase initiale à 39,5 % (257 sur 651) sous assistance, odds ratio ajusté 1,43 (1,11-1,84). Après retrait, il s'établit à 36,3 % (173 sur 476), et la différence avec la phase initiale n'est plus significative, odds ratio 1,03 (0,79-1,34). Chez les six expérimentés, aucun effet significatif de l'assistance sur le critère, dans aucune des trois phases. Les auteurs lisent l'absence de réduction post-exposition comme un argument contre l'érosion des compétences.

Ce point est décisif et il faut le formuler sans confort. *La contradiction subsiste lorsqu'on restreint la comparaison aux seuls opérateurs expérimentés. Elle ne peut donc pas être imputée au niveau d'expérience.* Dix-neuf experts polonais se dégradent, six experts scandinaves ne bougent pas. Les explications candidates sont nombreuses et aucune n'est tranchée : le critère n'est pas le même, détection d'adénomes contre détection de polypes d'au moins cinq millimètres ; le design non plus, rétrospectif avant-après contre prospectif à trois phases avec retrait explicite ; la puissance du sous-groupe expérimenté est faible, six opérateurs ; la structure et la durée de l'exposition diffèrent ; le cas-mix et le contexte organisationnel des centres aussi ; les ajustements retenus ne sont pas identiques ; et la définition même du deskilling n'est pas la même d'un protocole à l'autre. Les auteurs de l'essai prospectif relèvent d'ailleurs que leur groupe expérimenté était sensiblement plus âgé que le groupe non expérimenté, et suggèrent qu'un biais de conservatisme dans l'interaction humain-machine pourrait jouer.

Troisième niveau, l'état du champ.

Une revue systématique publiée en juillet 2026 a interrogé PubMed, Scopus, Web of Science et Google Scholar, identifié 11 269 références et retenu 29 études selon PRISMA 2020 (Al-Anezi F. M., « Generative Artificial Intelligence in Healthcare: Automation Bias, Deskilling, and Cognitive Implications », *J Healthc Leadersh* 2026;18:590498). Le biais d'automatisation y apparaît dans dix études, la préoccupation de déqualification dans neuf, l'atteinte au raisonnement diagnostique dans neuf également. Les auteurs décrivent une littérature à dominante observationnelle, expérimentale sur simulation ou purement conceptuelle, et trop hétérogène pour une synthèse quantitative. Autrement dit, la préoccupation est massivement partagée et l'appareil de preuve ne l'est pas.

Quatrième niveau, celui qui manque.

Établir qu'aucune mesure longitudinale de la vitesse de dépréciation et de la vitesse de reconstitution n'existe suppose une recherche négative dédiée et reproductible, avec bases interrogées, chaînes de recherche, date de coupure, critères d'inclusion et journal de screening. Elle n'a pas été conduite ici. Tant qu'elle ne l'est pas, ce texte affirme seulement que de telles mesures n'ont pas été identifiées, ce qui n'est pas la même chose. *Transformer une absence trouvée en absence démontrée est une spécialité académique dont l'efficacité s'arrête au premier rapporteur sérieux*.

Reste ce que les quatre niveaux établissent ensemble, et qui est plus solide que n'importe lequel d'entre eux pris isolément. La question de savoir si l'usage d'un système modifie la capacité de celui qui le supervise est aujourd'hui posée sans critère partagé, sans stratification convenue, sans horizon d'observation commun et sans définition stable de ce qu'on cherche. On ne peut pas trancher, non par manque de données mais par absence d'instrument : une discipline dotée d'indicateurs de qualité parmi les mieux normalisés de toute la médecine n'arrive pas à décider si son outil d'assistance déprécie ses praticiens. **Le problème intéressant n'est plus de savoir si la déqualification existe. C'est que nous n'avons ni estimand, ni protocole, ni horizon communs permettant de décider quand elle existe, chez qui, à quelle vitesse et avec quelle réversibilité.**

Le mécanisme, lui, est théorisé depuis quarante-trois ans. Lisanne Bainbridge, dans « Ironies of Automation » (*Automatica*, 19(6), 1983), établissait que l'automatisation retire à l'opérateur la conduite routinière, laquelle entretient la compétence requise pour les cas exceptionnels, lesquels sont ce que l'automatisation lui laisse. La délégation cognitive n'a rien de nouveau : diagnostic assisté, contrôle aérien, systèmes experts et contrôle de procédés déchargent des tâches cognitives depuis des décennies. *Avec l'IA générative, le paradoxe de Bainbridge s'étend à grande échelle à des tâches linguistiques, analytiques et normatives dont la vérité de référence est souvent incomplète, coûteuse ou absente.* Changement de terrain plutôt que de nature, et le terrain nouveau est celui où la mesure est la plus difficile.

Deux travaux récents sont cohérents avec cette lecture sans la démontrer. Lee et ses coauteurs (Microsoft Research et Carnegie Mellon, *CHI '25*, doi:10.1145/3706598.3713778) ont interrogé 319 travailleurs du savoir sur 936 usages réels : plus la confiance dans le système est élevée, moins la pensée critique est mise en œuvre, l'effet inverse valant pour la confiance en soi. L'étude est déclarative et mesure des perceptions, pas des performances. Kosmyna et ses coauteurs (arXiv:2506.08872, préimpression non revue, n = 54) rapportent la connectivité EEG la plus faible dans le groupe assisté et une incapacité des participants à citer leur propre production ; une critique méthodologique documentée en conteste la taille d'échantillon, la

reproductibilité et le traitement du signal (Stanković et al., arXiv:2601.00856), et le premier ne se cite pas sans la seconde.

Lin et ses coauteurs ont formulé l'observation générale plus tôt et plus nettement que je ne l'aurais fait (arXiv:2602.00854, janvier 2026) : la performance assistée s'améliore pendant que les modèles internes de l'utilisateur se dégradent, ce qu'ils nomment *Capability-Comprehension Gap*, et ils énoncent la nécessité d'un seuil minimal de compréhension en deçà duquel la supervision n'est plus qu'une procédure. Ils posent la question du seuil sans la trancher empiriquement, et ne traitent ni de la dynamique qui y conduit, ni de l'échelle organisationnelle, ni de l'instrumentation réglementaire. C'est l'espace que ce texte occupe.

3. Le modèle

La lecture naïve du résultat polonais consiste à dire que l'IA rend les gens moins bons. C'est vrai, banal et inexploitable. Pour en faire un objet de gouvernance, il faut cesser de traiter la compétence de contrôle comme un bloc.

Quatre capacités distinctes se cachent derrière le mot. *La production*, réaliser la tâche soi-même. *La vérification*, établir si une sortie est correcte par confrontation à un référentiel. *La récupération*, reprendre la main quand le système défaille, sous contrainte de temps. *L'escalade*, reconnaître que l'on ne sait pas et savoir vers qui basculer. Elles ne se déprécient ni au même rythme ni sous les mêmes conditions, et ne sont pas mobilisées par les mêmes régimes :

$$H_{nominal} = f(C_V, , C_E, A\ T)\ H_{dégradé} = f(C_R, , C_P, , C_E\ A\ T)$$

où A est l'autorité formelle et T le temps disponible pour intervenir. La capacité de production ne conditionne pas le contrôle courant : on peut contredire sans savoir refaire, par détection d'incohérences internes, calibration de l'incertitude, identification des cas hors distribution, recours à un second système indépendant ou retour aux sources primaires. *Savoir refaire et savoir contredire ne sont pas la même compétence.* Mais la capacité de production redevient critique en régime dégradé, quand il faut non plus juger une sortie mais produire soi-même un substitut acceptable. L'escalade figure dans les deux régimes, et pour des raisons différentes : reconnaître en fonctionnement normal qu'un cas dépasse sa compétence, reconnaître en mode dégradé qu'on ne reprendra pas la main seul. C'est la distinction classique entre fonctionnement nominal et mode dégradé, et elle lève une tension que la version précédente de ce texte laissait ouverte : on peut à la fois exclure la production du contrôle courant et vouloir la mesurer comme indicateur de résilience.

Dans « Le contrôle humain n'est pas une présence », j'avançais que la supervision n'est pas un état mais un débit. La compétence n'est qu'un facteur de ce débit, et TT est celui que les organisations compriment en premier.

Ce qui gouverne la dégradation n'est pas la nature de l'outil mais la disponibilité d'un référentiel externe à la chose vérifiée. L'automatisation classique déchargeait le geste en laissant intact l'instrument de mesure : le pilote automatique dévie, l'altimètre le dit. Il n'y a pas d'un côté les outils dotés d'altimètre et de l'autre l'IA générative qui en serait privée. Il y a un gradient, et il suit la tâche, pas la technologie.

Tâche	**Référentiel externe**	**Exposition à la substitution**
Calcul numérique	fort	faible
Code testable	fort	faible à moyen
Extraction documentaire	assez fort	moyen
Diagnostic clinique	partiel	moyen à fort
Analyse stratégique	faible	fort
Raisonnement normatif, arbitrage	faible	fort

Ce tableau décrit une exposition, pas un risque, et son ordinalité est illustrative, non mesurée. La nuance est décisive. *Le référentiel disponible n'est pas le référentiel consulté*. Une tâche fortement vérifiable produit du deskilling massif si personne ne consulte jamais le référentiel, et le développeur qui accepte une suggestion sans exécuter le test en est l'exemple quotidien. Inversement, une tâche faiblement vérifiable peut rester saine si l'humain travaille contre le système plutôt qu'avec lui. Le risque dépend donc de quatre termes et non d'un seul : le taux de délégation, la vérifiabilité externe de la tâche, l'engagement effectif de vérification, et la qualité du retour reçu.

Le taux de délégation lui-même recouvre des régimes opposés : *substitution* quand le système fait à la place, *augmentation* quand il étend l'espace de raisonnement, *tutorat* quand il produit un retour qui accroît la compétence, ce qui arrive et doit être dit, *complaisance* quand l'engagement se réduit sans que rien ne soit appris. Deux organisations utilisant le même modèle dans la même proportion peuvent produire des trajectoires opposées. La variable déterminante n'est pas le volume d'usage mais l'architecture de l'interaction, ce qui fait apparaître un objet que personne ne régule : *le design de la délégation*.

Sous cette réserve, la dynamique s'écrit. Soit C la capacité indépendante d'un opérateur, bornée par un plafond C_{max}, soit δ la part des cas traités sans engagement autonome, λ le taux de dépréciation par non-pratique et μ le taux d'acquisition par pratique :

$$\frac{dC}{dt} = -\lambda\,\delta\,C \;+\; \mu\,(1-\delta)\,(C_{max} - C)$$

La compétence devient ce qu'elle est : un stock borné, entretenu par l'exercice et érodé par la non-pratique. Le modèle reste un jouet sans valeur prédictive, mais il rend visible ce que le raisonnement réglementaire suppose, à savoir que λ vaut zéro. Il fait surtout apparaître la question que je ne posais pas : non pas si la compétence diminue, mais à quelle vitesse elle se reconstitue rapportée à la vitesse à laquelle elle se déprécie.

- Si $\tau_{recovery} \ll \tau_{decay}$, le problème est un problème de préparation avant intervention, sérieux mais gérable.
- Si $\tau_{recovery} \gg \tau_{decay}$, il y a hystérésis, et c'est seulement alors qu'on est fondé à parler de dette.

Aucune mesure longitudinale de ces deux constantes n'a été identifiée, ni dans la revue systématique citée plus haut ni dans les protocoles des deux essais qui s'opposent. C'est le trou empirique central de ce dossier, plus important que le chiffre polonais, et le confirmer exige la recherche négative dédiée annoncée en section 2.

Reste que δ n'est pas exogène, et c'est ce qui achève de faire du dispositif un système dynamique plutôt qu'une fonction de décroissance. Le taux de délégation dépend lui-même de la compétence, de la confiance, de la charge de travail et de la performance perçue du modèle. Un opérateur moins compétent délègue davantage, ce qui le rend moins compétent. Un expert délègue aussi davantage, mais pour une raison inverse, parce qu'il sait mieux reconnaître ce qui peut l'être. La même statistique d'usage recouvre donc une boucle de rétroaction positive et une allocation rationnelle, et rien dans un tableau de bord ne les distingue.

À quoi s'ajoute une seconde boucle, rarement formulée : dans certains déploiements, et vraisemblablement de manière croissante, les validations, infirmations et corrections humaines servent ensuite à entraîner, router, ajuster ou évaluer le système. Ailleurs, ce retour est simplement journalisé sans rien alimenter, et la boucle ne se referme pas. Il s'agit donc d'une classe d'architectures, pas d'une propriété universelle.

$$IA_t \to \delta_t \to C_{t+1} \to Supervision_{t+1} \to \acute{R}esultats_{t+1} \text{ avec } C_t \to \delta_t \text{ et } Supervision_t \to IA_{t+1}$$

Là où elle se referme, le système modifie progressivement le juge qui contribue ensuite à produire les données par lesquelles sont évaluées et la qualité du système et celle de sa

supervision. Si le contrôleur devient complaisant, le signal se dégrade sans que rien ne le signale. Aucune des grandeurs observables n'est alors identifiable isolément : un taux d'infirmation est produit conjointement par la qualité du modèle, la capacité humaine, la confiance, l'interface et la charge de travail, et rien ne permet de l'attribuer à l'un plutôt qu'à l'autre. Ce n'est plus du deskilling. C'est une co-évolution du contrôleur et du contrôlé, dans laquelle *la mesure elle-même devient endogène*.

4. Le changement d'échelle

L'unité d'analyse pertinente n'est pas l'opérateur mais l'organisation, et la compétence collective n'est pas la somme des compétences individuelles : $C_{org} \neq \sum_i C_i$.

Une organisation peut conserver des experts parfaitement compétents et perdre sa capacité de contrôle parce qu'ils sont trop peu nombreux, plus dans la boucle, sollicités trop tard, ou présents ailleurs qu'au point de décision.

Surtout, elle dépend de deux stocks et non d'un seul : les experts en place, et le mécanisme qui produit les suivants. Ce mécanisme passe par les tâches que l'organisation vient de déléguer. Le senior reste compétent cinq ans de plus. Le junior n'a jamais eu à apprendre ce que le modèle fait désormais à sa place, et ne deviendra donc pas le senior qu'il aurait été. *Une organisation peut ainsi satisfaire aujourd'hui son seuil de compétence tout en ayant déjà perdu la capacité de le satisfaire demain*. Le stock est conforme ; le flux qui le reconstitue ne l'est plus.

La signature de ce phénomène est celle d'une dette technique invisible. La performance nominale reste constante, parfois s'améliore, et rien n'apparaît dans les indicateurs tant que le composant ancien fonctionne. Ce qui décroît silencieusement est la résilience, c'est-à-dire la capacité à absorber une défaillance. En vocabulaire de fiabilité : la compétence humaine devient une capacité non redondante dont le temps moyen de rétablissement augmente pendant que les opérations nominales restent excellentes, et personne ne surveille le temps moyen de rétablissement d'un savoir-faire. Quand les seniors partent, la compétence organisationnelle ne décroît pas régulièrement, elle franchit une falaise. *Une dépréciation linéaire au niveau individuel produit une discontinuité au niveau collectif, et c'est la discontinuité qui fait l'accident*.

C'est ici que la métaphore comptable doit être tenue avec rigueur, parce que le vocabulaire courant confond deux objets. *Un actif se déprécie ; une dette s'accumule*. La compétence est le stock, la substitution cognitive est son mécanisme de dépréciation, et la dette n'est ni l'un ni l'autre : c'est un écart, entre la capacité requise et la capacité disponible.

$$D_t = \max\left(0,\ C_{requise,\,t} - C_{disponible,\,t}\right)$$

Écrite ainsi, elle rappelle qu'elle possède deux sources et non une seule. Elle croît quand la capacité disponible baisse, ce que décrit la section 3. Elle croît aussi quand la capacité requise augmente, à compétence rigoureusement inchangée, par exemple lorsqu'un système est étendu à des cas plus critiques ou à un périmètre plus large. *Une organisation peut s'endetter sans se déprécier.* Le vocabulaire se stabilise alors : le test de performance non assistée devient un test d'impairment, et la pratique récurrente une dépense de maintenance. Reste la question que la section 6 doit trancher, et qui est la seule qui compte : combien faut-il, au juste.

Un dernier mécanisme achève le tableau et il est le plus contre-intuitif. Plus le système est performant, moins l'humain a de raisons de le contredire. Moins il le contredit, moins il exerce son jugement indépendant. Moins ce jugement est exercé, moins il est disponible pour le cas rare sur lequel le système échoue. Or c'est exactement ce cas rare qui motive l'exigence de supervision. **À mesure que l'IA s'améliore, le contrôle humain devient moins souvent nécessaire et plus difficile précisément au moment où il devient nécessaire.**

5. La lacune réglementaire

L'article 14 du règlement (UE) 2024/1689 exige que les systèmes à haut risque soient conçus pour être effectivement supervisés par des personnes physiques. Son paragraphe 4 énumère ce que ces personnes doivent pouvoir faire : comprendre les capacités et les limites du système, interpréter correctement sa sortie, décider de ne pas l'utiliser, l'écarter, la renverser, interrompre le système. Le point (b) enjoint de *rester conscient* de la tendance à s'en remettre automatiquement à la sortie, désignée en toutes lettres comme biais d'automatisation, seul endroit du règlement où un phénomène cognitif précis est nommé. L'article 26(2) exige du déployeur qu'il confie la supervision à des personnes disposant de la compétence, de la formation et de l'autorité nécessaires.

Le dispositif est cohérent et repose sur une hypothèse non énoncée. *Rester conscient d'un biais n'est pas conserver la capacité de le corriger.*

La lacune doit être formulée avec précision, car sa version large serait fausse. Soutenir que le règlement n'exige pas le maintien de la compétence serait imprudent : un contradicteur reconstruirait fonctionnellement une telle obligation à partir des devoirs du déployeur, du système de management, de la surveillance après commercialisation ou de normes sectorielles, et il aurait de bons arguments. **Ce qui manque n'est pas une norme, c'est un instrument : le règlement exige une supervision compétente sans définir de métrique de récence de cette compétence, de seuil minimal de performance indépendante, ni de test périodique établissant que la capacité de**

supervision demeure recouvrable. La norme ISO/IEC 42001 ne comble pas l'écart : sa clause 7.2 reprend l'architecture générique des systèmes de management, déterminer les compétences requises, former, évaluer l'efficacité de la formation, conserver les preuves. Ce que l'auditeur y examine, ce sont des dossiers de qualification et des relevés de formation ; rien dans la clause n'appelle une mesure de la performance de l'opérateur privé de l'outil. Je n'en conclus pas qu'une telle mesure n'existe nulle part, ce qui demanderait la recherche négative décrite plus haut, mais que le texte qui sert de référence internationale pour un système de management de l'IA ne la demande pas. *La conformité atteste d'un titre ; elle n'atteste pas d'une aptitude actuelle.*

La discipline directement concernée par le chiffre polonais, elle, n'a pas attendu. La Société européenne d'endoscopie digestive a publié en 2026 une position issue d'un consensus Delphi formel, définissant un curriculum pour l'acquisition et le maintien de la compétence nécessaire à l'usage de l'IA en endoscopie (Mori Y., Kopylov U., Sinonquel P. et al., *Endoscopy* 2026;58:202-210). Ses recommandations disent où va le consensus : sécuriser d'abord la compétence en endoscopie standard, acquérir les fondamentaux de l'IA, reconnaître et atténuer les biais cognitifs de l'interaction humain-machine, éviter la surconfiance dans la décision clinique, et surveiller en continu les indicateurs de qualité tout au long de l'intégration du système. *Le secteur a produit en dix-huit mois ce que le socle horizontal ne contient pas.* C'est un contre-exemple à la version large de ma critique et une confirmation de sa version étroite : la compétence de supervision est bien un objet régulable, la démonstration en est faite, et elle n'est faite que là où une communauté professionnelle s'en est saisie discipline par discipline.

Un secteur a d'ailleurs rencontré le même problème quarante ans plus tôt et l'a résolu dans l'ordre inverse. L'aéronautique a d'abord tenu la licence pour une aptitude, puis a constaté que non. Le paragraphe 61.57 du titre 14 du *Code of Federal Regulations* interdit à un pilote d'emporter des passagers s'il n'a pas effectué trois décollages et trois atterrissages dans les quatre-vingt-dix jours précédents, seul aux commandes. Ce n'est pas une exigence de diplôme, c'est une exigence de pratique récente et non assistée, opposable, datée, vérifiable. En janvier 2013, la *Federal Aviation Administration* a publié le SAFO 13002, qui énonce que l'usage continu des systèmes de pilotage automatique ne renforce pas les connaissances et compétences du pilote en pilotage manuel et peut dégrader sa capacité à récupérer rapidement l'aéronef depuis un état non désiré. En mai 2017, jugeant apparemment que le message n'était pas passé, la même administration a republié le texte sous le titre SAFO 17007 en ajoutant le mot « proficiency ».

L'aviation dispose donc d'une exigence de récence, d'une doctrine explicite sur la dégradation par automatisation, et d'un dispositif périodique qui contrôle la performance non assistée. L'endoscopie digestive dispose d'un curriculum de maintien de compétence adopté par consensus. La gouvernance horizontale de l'IA dispose d'une obligation de rester conscient.

Ce que l'aéronautique a fini par admettre est que le contrôle n'est pas extérieur au système contrôlé. Le droit de l'IA raisonne en architecture statique de responsabilités : il attribue des rôles et vérifie leur existence à un instant donné. Il faudrait qu'il raisonne en système dynamique à boucle fermée, où la capacité de contrôle est un état endogène. Cela corrige une position antérieure. Sous le nom d'*Authority-Influence Gap*, j'avais distingué l'autorité formelle du contrôle cognitif effectif, distinction que d'autres formulent comme l'écart entre supervision nominale et supervision effective (arXiv:2604.00081). Cet écart n'est pas un état constaté, c'est une trajectoire : l'autorité est stable par construction juridique, la capacité de l'exercer suit la dynamique de la section 3, et l'écart s'ouvre mécaniquement. Quant au contrôle humain, il figure au bilan de conformité comme un actif permanent, non amortissable, jamais soumis à test de dépréciation, et dont l'usage est exactement ce que le déploiement a pour objet de réduire.

Le règlement (UE) 2026/1744, entré en vigueur le 27 juillet 2026, a reporté au 2 décembre 2027 l'application des principales obligations pour les systèmes à haut risque autonomes de l'annexe III, sans modifier la substance de l'article 14. Dix-huit mois de délégation supplémentaire avant que l'actif ne soit inventorié.

6. L'instrumentation

Ce qui précède n'établit pas la thèse, il en établit la plausibilité et l'absence d'instrument. Quatre limites doivent être tenues. La base empirique est mince et divergente : deux essais sérieux concluent en sens opposé y compris sur des populations expérimentées, une revue systématique décrit une littérature hétérogène et largement observationnelle, et rien ne dit aujourd'hui lequel des deux résultats décrit le cas général. Un résultat obtenu sur un geste perceptuel ne se transporte pas mécaniquement au raisonnement analytique, et c'est une extrapolation nommée comme telle. Les constantes de dépréciation et de reconstitution sont inconnues, les horizons observés courts alors que la thèse porte sur un phénomène cumulatif. Enfin, la position défendue est plus facile à énoncer qu'à réfuter, ce qui est le reproche adressé au discours existentiel en ouverture : je ne prétends pas y échapper, je prétends fournir de quoi en sortir.

Il faut commencer par la question qu'aucune doctrine du deskilling ne pose jamais et qui conditionne tout le reste : combien de capacité indépendante faut-il conserver. Rien n'oblige une organisation à préserver cent pour cent de sa capacité autonome historique. Elle ne sait plus graver ses microprocesseurs, réparer ses ascenseurs ni calculer une paie à la main, et c'est parfaitement rationnel. *Toute dépréciation n'est pas une dette.* La dette n'apparaît qu'au franchissement d'un seuil :

$$C_{disponible} < C_{requise}\left(\acute{\text{r}}\text{esilience}\right)$$

Et six paramètres en déterminent conceptuellement le niveau, dont aucun n'est cognitif : criticité de la tâche, fréquence des défaillances du système, détectabilité de ces défaillances, temps nécessaire à la récupération, disponibilité d'alternatives indépendantes, réversibilité du dommage. Ce qu'il faut définir n'est donc pas un niveau de compétence humaine, c'est une capacité indépendante minimale de résilience : la capacité que l'organisation doit conserver pour détecter une défaillance, interrompre le système, maintenir un état sûr et restaurer un service acceptable. C'est l'analogue exact du service minimum viable en résilience informatique, transposé à la compétence.

Ce déplacement lève la plupart des objections adressées aux doctrines du maintien de compétence : il n'oblige à conserver ni toutes les anciennes compétences ni, chez chaque superviseur, la capacité de tout refaire, puisque la compétence peut être distribuée et l'escalade organisée, et le seuil est proportionné au risque plutôt qu'uniforme. *L'objectif n'est pas de préserver l'autonomie humaine maximale, c'est de préserver l'autonomie minimale nécessaire à la maîtrise du risque.* Il se distingue enfin du seuil de compréhension proposé par Lin et ses coauteurs, qui est épistémique et porte sur la contestabilité d'une sortie : le seuil de résilience est opérationnel et porte sur la capacité à tenir le système en état sûr sans lui. Les deux sont nécessaires et ne se substituent pas.

Vient ensuite ce qu'il faut mesurer.

- Premier renoncement : l'indicateur unique. Le taux d'infirmation semble le candidat évident, et il ne vaut rien seul : une baisse peut signifier une amélioration du modèle, une dégradation du contrôleur, une confiance excessive, une charge de travail accrue, un coût organisationnel de la contradiction, une interface qui décourage le désaveu, une sélection de cas plus simples ou une pression temporelle. Un indicateur qui admet huit explications concurrentes n'est pas un indicateur. Ce qu'il faut construire est un panel, c'est-à-dire une observabilité du contrôle humain : performance non assistée en aveugle, taux d'infirmation, part des infirmations correctes, taux de faux désaveux, calibration entre confiance déclarée et exactitude, délai avant détection, capacité de justification indépendante, et performance après interruption prolongée de l'assistance, qui est le seul moyen d'estimer $\tau_{recovery}$.
- Deuxième point : dédoubler le protocole. Mesurer l'opérateur sans le système répond à la question de savoir s'il sait encore faire, ce qui n'est pas la question réglementaire. Savoir s'il sait superviser en conditions réelles est un autre estimand. Il faut donc un *test de compétence autonome*, qui mesure le stock, et un *test d'efficacité de supervision*, qui injecte dans un environnement contrôlé une proportion définie de sorties erronées et mesure taux de détection, taux de correction et délai avant détection. Une forme de red teaming appliquée non au modèle mais au superviseur.

Encore faut-il injecter les bonnes erreurs, faute de quoi le test mesure le trivial. Une hallucination grossière n'apprend rien. Une réponse fausse, parfaitement argumentée, correctement sourcée sauf sur une prémisse discrète, apprend beaucoup. Le protocole doit donc stratifier :

$$E = \{Grossière, Plausible, Argumentée\ d'autorité, omission, cadrage, corrélée\}$$

Et produire non pas un score mais une probabilité de détection conditionnée à la difficulté et au type d'erreur. Une omission et une contre-vérité factuelle peuvent présenter la même difficulté nominale et solliciter des compétences différentes : ce qu'on obtient n'est donc pas une courbe mais une surface de performance de supervision, ce qui est cohérent avec le refus de l'indicateur unique. On obtient alors ce qui manque aujourd'hui à la conformité : non pas l'attestation qu'un humain était présent, mais la caractérisation de ce qu'il détecte et de ce qui lui échappe. C'est un programme expérimental et non une clause contractuelle, et c'est précisément ce qui le rend utile.

La transposition aéronautique ne doit pas pour autant être littérale. Exiger un volume de cas traités sans assistance serait prématuré : dix cas triviaux ne valent pas un cas critique, et les événements qu'il faut savoir détecter sont souvent trop rares pour être rencontrés naturellement. Ce que l'aviation enseigne n'est pas un quota brut mais une exigence de proficiency récurrente, complétée par la simulation lorsque l'exposition naturelle est insuffisante. Dont acte : cas synthétiques, scénarios rares et erreurs injectées deviennent l'instrument de maintien de la compétence humaine face à l'IA. Les populations simulées, longtemps discutées comme substitut de données, trouvent ici une fonction moins contestable.

Ce qui impose une condition que tout le raisonnement précédent rend obligatoire. Si le fournisseur du système produit également les cas de test, les erreurs injectées et le barème, on reconstitue exactement le couplage que l'on prétend mesurer : le dispositif d'évaluation devient un organe du système évalué. *L'indépendance du test à l'égard du système testé est une propriété de conception, pas une bonne pratique*. Elle s'obtient par des modèles ou des fournisseurs distincts, des jeux de cas validés extérieurement, une vérité de référence externe, ou une génération synthétique suivie d'une validation indépendante. À défaut, on demande au modèle de rédiger l'examen qui vérifiera si l'humain sait repérer ses erreurs, ce qui constituerait un protocole d'évaluation assez créatif.

Il faut enfin dire ce que cela coûte, puisque c'est le point sur lequel un directeur financier tranchera. Une capacité indépendante minimale n'est pas gratuite : cas traités sans assistance, simulations périodiques, doubles lectures, exercices adverses, temps d'expert, capacité de secours. Son maintien constitue une dépense récurrente de résilience, et cette dépense appartient au coût total de l'automatisation, non aux frais généraux de la conformité.

Il reste à ranger les objets dans le bon ordre. Extinction, dépossession et dépendance cognitive ne sont pas trois risques comparables : le premier est une conséquence, le deuxième un état, le troisième un mécanisme, et les aligner reviendrait à commettre l'erreur reprochée en ouverture. La forme correcte en sépare trois niveaux. Des mécanismes, dont la dépréciation cognitive, le biais d'automatisation et la contamination du signal de supervision. Un état intermédiaire, la perte de supervision effective. Des conséquences, des erreurs non détectées jusqu'à la catastrophe. L'extinction occupe la dernière case de la troisième colonne, avec une probabilité indécidable. La substitution cognitive occupe la première case de la première, avec un mécanisme documenté depuis 1983, deux essais qui se contredisent faute de protocole commun, une revue systématique qui constate l'hétérogénéité du champ, un précédent réglementaire dans un secteur voisin, et un curriculum professionnel qui prouve que la chose est régulable. Travailler sur la première colonne n'est pas renoncer à la troisième, c'est la seule façon d'y arriver.

Le contrôle humain n'est pas une garantie que l'on inscrit au dossier. C'est une variable d'état du système qu'il est censé contrôler, et la question n'est pas de savoir si elle baisse. La question est de savoir quel niveau doit être tenu, et de disposer d'un instrument pour établir qu'il l'est. Nous garderons l'autorité. **Il n'est pas nécessaire que la machine prenne le contrôle si nous cessons d'entretenir ce qui nous permettrait de l'exercer.**